DOCUMENT RESUME

ED 314 450 TM 014 265

AUTHOR Davidson, Betty M.; Giroir, Mary M.
TITLE Comments on Better Uses of and Alternatives to Significance Testing.
PUB DATE 9 Nov 89
NOTE 18p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Reliability; Research Design; Researchers; *Scientific Methodology; Statistical Analysis; *Statistical Significance
IDENTIFIERS *Significance Testing

ABSTRACT

Controversy over the proper place of significance testing within scientific methodology has continued for some time. The suggestion that effect sizes are more important than whether results are significant is presented. Effect size can be defined as an estimate of how much of the dependent variable is accounted for by the independent variables. Interpretations of statistical significance can be seriously incorrect when the researcher underinterprets an outcome with a large effect size that is nonsignificant or overinterprets an outcome that involves a small effect size but which is statistically significant. These problems can be avoided if the researcher includes effect size in result interpretation. It has been stated that statistical significance was never intended to take the place of replication in research. Researchers must begin drawing conclusions based on effect sizes and not statistical significance alone; and the replicability and reliability of results must be recognized, analyzed, and interpreted. Two tables illustrate effect sizes. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314450

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BETTY M. DAVIDSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

COMMENTS ON BETTER USES OF AND ALTERNATIVES TO SIGNIFICANCE TESTING

Betty M. Davidson Mary M. Giroir

University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 9, 1989.

TM014265

BEST COPY AVAILABLE

ABSTRACT

Since almost the beginning of the century a continuing controversy has brewed over statistical significance testing and its proper place within the scientific method. It is suggested in this paper that effect sizes are much more important than whether results are "significant." Small data sets are presented to illustrate how effect sizes can be employed to good result, and to evaluate the influence of sample size on significance tests.

Since the promulgation of Sir Ronald Fisher's logic of null hypothesis testing in the 1920's educational researchers have felt compelled to employ significance testing to interpret data collected from various populations. A continuing controversy has brewed over statistical significance testing and its proper place within the scientific method. (Carver, 1978; Keasler, 1988; Kupfersmid, 1988; Thompson, 1989b). Once dominating the literature, significance testing still is included in many of the educational and psychological journals; some authors only submit articles in which the results are statistically significant (Carver, 1978; Greenwald, 1975; Keaster, 1988; Thompson, 1987). But as Meehl (1978, pp. 817; 823) states emphatically:

> I suggest to you that Sir Ronald (Fisher) has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substative theories in the soft (i.e., social science) areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology.... I am not making some nit-picking statistician's correction. I am saying that the whole business

> is so radically defective as to be scientifically almost pointless.

Cronback (1975, p. 124) asserts that "the time has arrived to exorcise the null hypothesis." The general purpose of the present paper is to analyze statistical significance testing and to explore alternative methods to better interpret data and thereby increase the knowledge base.

These improvements may be important. As Thompson (1987, p. i) states:

> Reliance on the traditional hypothesis testing model has led to a major bias against nonsignificant results and to misinterpretation of significant results. A finding of statistical significance does not mean that the null hypothesis is false, since there are many factors affecting statistical significance such as sample size and the measurement reliability of the data. Furthermore, statistical significance also does not permit evaluation of the importance of a finding.

Carver (1978, p. 378) argues that overemphasis on statistical significance in education research has resulted in a "corrupt form of scientific method." Shaver (1979, pp. 5-6) argues that: "The emphasis on statistics and the 'test of significance' procedure has resulted in a methodological orientation toward establishing generalizability that has been deleterious in its effects of

the scientific accumulation of knowledge in education." Carver (1978, p. 298) states, "Educational research would be better off if it stopped testing its results for statistical significance."

Thompson (1989a, p. 66) states, "Few statistical procedures have caused more confusion within the research community than statistical significance testing." Researchers must be reminded that obtaining statistical significance, by itself, does not make the results important. Thompson (1987, p.10) asserts, "Significance tests do not require as input into calculations declarations of the researcher's value system, and therefore cannot contain as output information about the importance of results."

INFLUENCE OF SAMPLE SIZE

Schneider and Darcy (1984, p. 575) identified seven factors which determine the outcome of significance tests:

1) Actual strength of impact
2) Number of cases used in the study
3) Variation among cases on relevant variables
4) The complexity of the analysis (degrees of freedom)
5) The appropriateness of the statistical measures and tests used
6) The hypothesis tested
7) The significance level chosen

Although significance testing is a function of these interrelated features, sample size is a basic influence on the results of significance testing (Keaster, 1988; Thompson, 1988a).

Not enough is written about the researcher's ability to control statistical significance by controlling sample size. As Kish (1959, p. 336) states, "The results of statistical 'tests of significance' are functions not only of the magnitude of the relationships studied but also of the numbers of sampling units used." Hays (1981, p. 293) states even more emphatically that, "Virtually any study can be made to show significant results if one uses enough subjects."

The decision to reject or not reject the null hypothesis must be interpreted in the context of the sample size. A significant result that would no longer be significant with a few fewer subjects is troublesome; a nonsignificant result that would be significant with a few more subjects certainly may be noteworthy. According to Thomspon (1988a, p. 147), "Thus, results may be more noteworthy when results are significant and sample size is small, or when results are not significant and sample size is large. This suggests that when authors interpret significance, they should consider sample size."

Table 1 illustrates significance testing with increasing sample sizes (55, 56, 57, 58, 59 60) but with a fixed

effect size. Note that the critical F at a fixed alpha of 0.05 gets smaller as degrees of freedom error increases with increasing sample size. Also, as the degrees of freedom error increases, the mean square error gets smaller, and the calculated F therefore gets larger. As illustrated, sample size does effect the "yes-no" hypothesis test decision.

INSERT TABLE 1 ABOUT HERE

IMPORTANCE OF EFFECT SIZE

Another intepretation aid emphasizes effect sizes. Effect size can be defined as an estimate of how much of the dependent variable is accounted for by the independent variables (Keaster, 1988). One simple effect size is the correlation ratio, which is computed by dividing the sum of squares for an effect by the total sum of squares. A correlation ratio equilvalent to 25% is considered quite large (Cohen, 1988).

Correlation ratio and eta-squared are synonymous terms. With only one predicator variable, r (yx) = R (yx) and r squared (yx) = R squared (yx). Regardless of the number of predicator variables (PV), R (y * x1, x2, ...) = r (y, YHAT) and R squared (y * x1, x2, ...) = r squared (y, YHAT). Squaring of r and R is necessary in order to compare the common variance. Eta-squared, r squared and R squared are in the same metric of common variance; they explain what

proportion of the variance or variables is shared by the predicator and dependent variables.

The correlation ratio, r, R, r squared and R squared, only measure a linear relationship between variables. Eta-squared is sensitive both to linear and non-linear relationships. Thus, eta-squared is always greater than or equal to r squared for any given data set. If two variables demonstrate a purely linear relationship, eta-squared and r squared will be equivalent. The correlation ratio or eta-squared is to indicate the portion of the dependent variable(s) explained by the independent variable(s). The effect size, correlation ratio and eta-squared, is not influenced by sample size and therefore results are less ambiguous when effect sizes are reported and interpreted.

When conducting research the goal is to explain as much of the dependent variable as possible (Keaster, 1988; Thompson, 1988a). As Yates (1951, pp. 32-33) notes:

> (The use of tests of significance) has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating... . The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the

> unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it.

Unfortunately, as Shaver (1980, p. 13) notes, "Making decisions about the size an effect must reach to be important is usually not easy in educational research, for a number of reasons. One is that current research paradigms have not conditioned us to think that way."

Interpretations of statistical significance can be seriously wrong when the researcher underinterprets an outcome with a large effect size that is nonsignificant, or overinterprets an outcome that involves a small effect size but which is statistically significant. This can be avoided by the researcher including effect size within the rubic of result interpretation. As Loftus and Loftus (1982, p. 499) suggest, "It is our judgment that accounting for variance is really much more meaningful than testing for significance."

As Table 2 indicates, the effect size remains constant even when the sample size varies. In a research study the null hypothesis could be rejected, i.e., be statistically significant, but a small effect size could exist. Even though the results would be statistically significant only a small portion of the dependent variable would be explained by the independent variable(s). The opposite is also true,

the research results could be not statistically significant, the null hypothesis would not be rejected, but a large effect size could exist. A greater portion of the dependent variable would be explained by the independent variable(s) even though the results were not statistically significant. Thus, effect size is a very important indicator that is not directly influenced by sample size.

INSERT TABLE 2 ABOUT HERE

As Thompson (1987, p. 23) suggests:

> The failure to utilize effect size estimates in interpretation may partially be due to the influence of the significance testing paradaigm which influences attention away from effect size. The failure to consult effect sizes also may be due to another paradigm problem involving researchers not recognizing that all analyses actually test association.

As McNamara (1974, p. 50) concurs, noting that "Rejection of the hypothesis of no difference between population means is tantamount to the assertion that the independent variables do have some statistical association with the criterion scores."

As Thompson (1988a, p. 147) notes, "using effect sizes in interpretation focuses interpretation on what the

researcher really cares about (i.e., the issue of noteworthiness of results)."

IMPORTANCE OF REPLICATION

Tukey (1969, p. 85) states that significance testing was never intended to take the place of replication in research. Replication of research designs using different populations allows the researcher to verify the validity of the study's results. Carver (1978, pp. 385-386) states,

> Whether or not the results are replicable or reliable actually depends upon whether the important variables can be controlled and manipulated in exactly the same way in order to give the same results. Certainly, there is no magic in the numbers collected and analyzed using the assumptions of the null hypothesis that allows us to infer anything about the probability that another researcher will be able to get a similar mean difference. Too often statistical significance is substituted for actual replicative evidence; too often statistical significance covers up an inferior research design.

Thompson (1987, April, p. 27) asserts, "Researchers have increasingly recognized the critical nature of replication as the utlimate test of scientific findings and some have argued that replicability should replace significance testing as part of a new logic of truth

testing." Carver (1978, p. 386) agrees, noting that "It is a fantasy to hold that statistical significance reflects the degree of confidence in the replicability or reliability of results."

Statistical significance can no longer be considered "the way" to interpret scientific results. Researchers must begin drawing conclusions based on effect sizes and not just statistical significance; replicability and reliability of results must also be recognized, analyzed and interpreted. Researchers must accept that the "fantasy" is over and that significance testing is no longer the cornerstone for interpreting social science research results.

REFERENCES

Carver, R. P. (1978). The case against significance testing. Harvard Educational Review, 48(3), 378-399.

Cronback, L. J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. Journal of Counsulting and Clinical Psychology, 47, 421-423.

Greenwald, A. G. (1975). Consequences of the prejudice against the null hypothesis. Psychological Bulletin, 82, 1-20.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart & Winston.

Keaster, R. D. (1988, November). Statistical significance testing: from routine to ritual. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-883.

Kupersmid, J. (1988). Improving what is published: a model in search of an editor. American Psychologist, 43, 635-642.

Loftus, G. R., & Loftus, E. F. (1982). Essence of statistics. Monterey, CA: Brooks/Cole.

McNamara, J. F. (1978). Practical significance and statistical models. Educational Administration Quarterly, 14, 48-63.

Meehl, P. (1978). Theoretical risk and tabular asterisk: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Scheider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8(4), 573-583.

Shaver, J. P. (1980, April). Readdressing the role of statistical tests of significance. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 191 904)

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 287 868)

Thompson, B. (1988a). A note about significance testing. Measurement and Evaluation in Counseloring and Development, 20(4), 146-148.

Thompson, B. (1988b, November). Common methodology mistakes

in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1989a). Asking "what if" questions about significance tests. Measurement and Evaluation in Counseling and Development, 20(4), 146-148.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseloring and Development, 22(4), 2-5.

Tukey, J. W. (1969). Analyzing data: sanctification or detective work? American Psychologist, 24, 83-91.

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.

TABLE 1

Statistical Significance at Various Sample Sizes
for a Moderate Fixed Effect Size

Source	SOS	Effect Size	df	MS	Fcalc	Fcrit	Dec
SOS exp	2.29	0.099	2	1.15	3.19	3.16	Rej
SOS unexp	20.70		57	0.36			
SOS tot	23.00		59				
SOS exp	2.29	0.099	2	1.15	3.11	3.16	Not
SOS unexp	20.70		56	0.37			Rej
SOS tot	23.00		58				
SOS exp	2.29	0.099	2	1.15	3.03	3.16	Not
SOS unexp	20.07		55	0.38			Rej
SOS tot	23.00		57				
SOS exp	2.29	0.099	2	1.15	3.03	3.17	Not
SOS unexp	20.70		54	0.38			Rej
SOS tot	23.00		56				
SOS exp	2.29	0.099	2	1.15	2.90	3.17	Not
SOS unexp	20.70		53	0.39			Rej
SOS tot	23.00		55				
SOS exp	2.29	0.099	2	1.15	2.88	3.18	Not
SOS unexp	20.07		52	0.40			Rej
SOS tot	23.00		54				

TABLE 2

Statistical Significance at Various Sample Sizes
for a Large Fixed Effect Size
(alpha = 0.05)

Source	SOS	Effect Size	df	MS	Fcalc	Fcrit	Dec
SOS exp	337.2	0.336	2	168.600	0.253	200.00	Not
SOS unexp	665.1		1	665.100			Rej
SOS tot	1002.3		3	334.100			
SOS exp	337.2	0.336	2	168.600	2.535	4.10	Not
SOS unexp	665.1		10	66.510			Rej
SOS tot	1002.3		12	83.525			
SOS exp	337.2	0.336	2	168.600	5.070	3.49	Rej
SOS unexp	665.1		20	33.255			
SOS tot	1002.3		22	45.559			
SOS exp	337.2	0.336	2	168.600	7.605	3.32	Rej
SOS unexp	665.1		30	22.170			
SOS tot	1002.3		32	31.322			

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22(4), 2-5.